Andmete sisselugemine

Andmete lugemine

Andmeid on võimalik sisse lugeda erinevas formaadis. Lihtsam viis on excel, vsc, txt jms failid

```
states=read.csv('data/states.csv')
class(states)
```

```
## [1] "data.frame"
```

```
dim(states)
```

```
## [1] 51 21
```

Muutujate kirjeldus:

```
str(states)
```

```
## 'data.frame':    51 obs. of  21 variables:
##  $ state  : Factor w/ 51 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ region : Factor w/ 4 levels "Midwest","N. East",..: 3 4 4 3 4 4 2 3 NA 3 ...
##  $ pop    : int  4041000 550000 3665000 2351000 29760000 3294000 3287000 666000 NA 12938000 ...
##  $ area   : int  52423 570374 113642 52075 155973 103730 4845 1955 NA 53997 ...
##  $ density: num  77.08 0.96 32.25 45.15 190.8 ...
##  $ metro  : num  67.4 41.1 79 40.1 95.7 ...
##  $ waste  : num  1.11 0.91 0.79 0.85 1.51 ...
##  $ energy : int  393 991 258 330 246 273 234 349 NA 237 ...
##  $ miles  : int  10500 7200 9700 8900 8700 8300 8000 9800 NA 8500 ...
##  $ toxic  : num  27.86 37.41 19.65 24.6 3.26 ...
##  $ green  : num  29.2 NA 18.4 26 15.6 ...
##  $ house  : int  30 0 13 25 50 36 64 69 NA 45 ...
##  $ senate : int  10 20 33 37 47 58 87 83 NA 47 ...
##  $ csat   : int  991 920 932 1005 897 959 897 892 840 882 ...
##  $ vsat   : int  476 439 442 482 415 453 429 428 405 416 ...
##  $ msat   : int  515 481 490 523 482 506 468 464 435 466 ...
##  $ percent: int  8 41 26 6 47 29 81 61 71 48 ...
##  $ expense: int  3627 8330 4309 3700 4491 5064 7602 5865 9259 5276 ...
##  $ income : int  27498 48254 32093 24643 41716 35123 48618 40641 35807 32027 ...
##  $ high   : num  66.9 86.6 78.7 66.3 76.2 ...
##  $ college: num  15.7 23 20.3 13.3 23.4 ...
```

Alati kontrolli, kas asjad on korras:

```
head(states)
```

```
##        state region      pop   area density metro waste energy miles toxic
## 1    Alabama  South  4041000  52423   77.08  67.4  1.11    393 10500 27.86
## 2     Alaska   West   550000 570374    0.96  41.1  0.91    991  7200 37.41
## 3    Arizona   West  3665000 113642   32.25  79.0  0.79    258  9700 19.65
## 4   Arkansas  South  2351000  52075   45.15  40.1  0.85    330  8900 24.60
## 5 California   West 29760000 155973  190.80  95.7  1.51    246  8700  3.26
## 6   Colorado   West  3294000 103730   31.76  81.5  0.73    273  8300  2.25
##   green house senate csat vsat msat percent expense income high college
## 1 29.25    30     10  991  476  515       8    3627  27498 66.9    15.7
## 2    NA     0     20  920  439  481      41    8330  48254 86.6    23.0
```

```
## 3 18.37    13     33  932  442  490      26    4309  32093 78.7    20.3
## 4 26.04    25     37 1005  482  523       6    3700  24643 66.3    13.3
## 5 15.65    50     47  897  415  482      47    4491  41716 76.2    23.4
## 6 21.89    36     58  959  453  506      29    5064  35123 84.4    27.0
```

```
tail(states)
```

```
##            state  region     pop  area density metro waste energy miles
## 46       Vermont N. East  563000  9249   60.87  23.4  0.69    232 10400
## 47      Virginia   South 6187000 39598  156.25  72.5  1.45    306  9700
## 48    Washington    West 4867000 66582   73.10  81.7  1.05    389  9200
## 49 West Virginia   South 1793000 24087   74.44  36.4  0.95    415  8600
## 50     Wisconsin Midwest 4892000 54314   90.07  67.4  0.70    288  9100
## 51       Wyoming    West  454000 97105    4.68  29.6  0.70    786 12800
##    toxic  green house senate csat vsat msat percent expense income high
## 46  1.81  15.17    85     94  890  424  466      68    6738  34717 80.8
## 47 12.87  18.72    33     54  890  424  466      60    4836  38838 75.2
## 48  8.51  16.51    52     64  913  433  480      49    5000  36338 83.8
## 49 21.30  51.14    48     57  926  441  485      17    4911  24233 66.0
## 50  9.20  20.58    47     57 1023  481  542      11    5871  34309 78.6
## 51 25.51 114.40     0     10  980  466  514      13    5723  31576 83.0
##    college
## 46    24.3
## 47    24.5
## 48    22.9
## 49    12.3
## 50    17.7
## 51    18.8
```

Kiire ülevaade muutujatest:

```
summary(states)
```

```
##         state        region        pop               area
##  Alabama   : 1   Midwest:12   Min.   :  454000   Min.   :  1045
##  Alaska    : 1   N. East: 9   1st Qu.: 1299750   1st Qu.: 36802
##  Arizona   : 1   South  :16   Median : 3390500   Median : 54156
##  Arkansas  : 1   West   :13   Mean   : 4962040   Mean   : 70759
##  California: 1   NA's   : 1   3rd Qu.: 5898000   3rd Qu.: 81272
##  Colorado  : 1                Max.   :29760000   Max.   :570374
##  (Other)   :45                NA's   :1          NA's   :1
##     density            metro            waste            energy
##  Min.   :   0.96   Min.   : 20.40   Min.   :0.5400   Min.   :200.0
##  1st Qu.:  31.88   1st Qu.: 46.98   1st Qu.:0.8225   1st Qu.:285.0
##  Median :  75.76   Median : 67.55   Median :0.9600   Median :320.0
##  Mean   : 166.04   Mean   : 64.07   Mean   :0.9888   Mean   :354.5
##  3rd Qu.: 170.29   3rd Qu.: 81.58   3rd Qu.:1.1450   3rd Qu.:371.5
##  Max.   :1041.92   Max.   :100.00   Max.   :1.5100   Max.   :991.0
##  NA's   :1         NA's   :1        NA's   :1        NA's   :1
##      miles           toxic             green            house
##  Min.   : 5900   Min.   :  0.770   Min.   : 11.76   Min.   : 0.00
##  1st Qu.: 8500   1st Qu.:  6.737   1st Qu.: 16.98   1st Qu.:31.00
##  Median : 9100   Median : 11.705   Median : 21.38   Median :44.50
##  Mean   : 9046   Mean   : 17.606   Mean   : 25.11   Mean   :44.82
##  3rd Qu.: 9700   3rd Qu.: 21.488   3rd Qu.: 26.34   3rd Qu.:59.25
##  Max.   :12800   Max.   :101.280   Max.   :114.40   Max.   :85.00
```

```
##  NA's   :1       NA's   :1         NA's   :3        NA's   :1
##      senate           csat             vsat            msat
##  Min.   :10.00   Min.   : 832.0   Min.   :395.0   Min.   :435.0
##  1st Qu.:27.00   1st Qu.: 888.0   1st Qu.:421.0   1st Qu.:467.0
##  Median :51.00   Median : 926.0   Median :441.0   Median :485.0
##  Mean   :49.78   Mean   : 944.1   Mean   :447.8   Mean   :496.3
##  3rd Qu.:67.00   3rd Qu.: 997.0   3rd Qu.:476.0   3rd Qu.:521.5
##  Max.   :97.00   Max.   :1093.0   Max.   :515.0   Max.   :578.0
##  NA's   :1
##     percent         expense          income           high
##  Min.   : 4.00   Min.   :2960   Min.   :23465   Min.   :64.30
##  1st Qu.:11.00   1st Qu.:4352   1st Qu.:29875   1st Qu.:73.50
##  Median :26.00   Median :5000   Median :33452   Median :76.70
##  Mean   :35.76   Mean   :5236   Mean   :33957   Mean   :76.26
##  3rd Qu.:60.50   3rd Qu.:5794   3rd Qu.:36920   3rd Qu.:80.10
##  Max.   :81.00   Max.   :9259   Max.   :48618   Max.   :86.60
##
##     college
##  Min.   :12.30
##  1st Qu.:17.30
##  Median :19.30
##  Mean   :20.02
##  3rd Qu.:22.90
##  Max.   :33.30
##
```

Sisselugemise juures on oluline:

```
  - mis on eraldusmärk (kui see erineb arvuti sinu vaikimise märgist,
  tuleb see ette anda argumendina näiteks: sep=";")
  - kas read ja veerud on korrektsed, csv-s võivad need lappama minna,
  kui csv on loodud mõnes teises operatsioonisüsteemis!
  - vael aitab read_csv2, kui read_csv-ga asjad untsu lähevad
```

Exceli sisselugemine on sarnane:

```
library(readxl)
hdi=read_excel("data/HDIdat.xls.xlsx")
```

```
head(hdi, 20)
```

```
## # A tibble: 20 x 25
##    `International ~ X__1  X__2  X__3  X__4  X__5  X__6  X__7  X__8  X__9
##    <chr>            <chr> <chr> <lgl> <chr> <lgl> <chr> <lgl> <chr> <lgl>
##  1 <NA>             <NA>  <NA>  NA    <NA>  NA    <NA>  NA    <NA>  NA
##  2 <NA>             <NA>  <NA>  NA    <NA>  NA    <NA>  NA    <NA>  NA
##  3 Accessed: 10/31~ <NA>  <NA>  NA    <NA>  NA    <NA>  NA    <NA>  NA
##  4 <NA>             <NA>  <NA>  NA    <NA>  NA    <NA>  NA    <NA>  NA
##  5 Human Developme~ <NA>  <NA>  NA    <NA>  NA    <NA>  NA    <NA>  NA
##  6 <NA>             <NA>  <NA>  NA    <NA>  NA    <NA>  NA    <NA>  NA
##  7 A composite ind~ <NA>  <NA>  NA    <NA>  NA    <NA>  NA    <NA>  NA
##  8 <NA>             <NA>  <NA>  NA    <NA>  NA    <NA>  NA    <NA>  NA
##  9 Source: HDRO ca~ <NA>  <NA>  NA    <NA>  NA    <NA>  NA    <NA>  NA
## 10 <NA>             <NA>  <NA>  NA    <NA>  NA    <NA>  NA    <NA>  NA
## 11 Data in the tab~ <NA>  <NA>  NA    <NA>  NA    <NA>  NA    <NA>  NA
```

```
## 12 <NA>             <NA>  <NA>  NA    <NA>  NA    <NA>  NA    <NA>  NA
## 13 HDI Rank         Coun~ 1980  NA    1985  NA    1990  NA    1995  NA
## 14 ..               Very~ 0.766 NA    0.786 NA    0.810 NA    0.834 NA
## 15 ..               High~ 0.614 NA    0.630 NA    0.648 NA    0.662 NA
## 16 ..               Medi~ 0.420 NA    0.450 NA    0.480 NA    0.517 NA
## 17 ..               Low ~ 0.316 NA    0.334 NA    0.347 NA    0.363 NA
## 18 1                Norw~ 0.796 NA    0.819 NA    0.844 NA    0.876 NA
## 19 2                Aust~ 0.850 NA    0.859 NA    0.873 NA    0.889 NA
## 20 3                Neth~ 0.792 NA    0.806 NA    0.835 NA    0.866 NA
## # ... with 15 more variables: X__10 <chr>, X__11 <lgl>, X__12 <chr>,
## #   X__13 <lgl>, X__14 <chr>, X__15 <lgl>, X__16 <chr>, X__17 <lgl>,
## #   X__18 <chr>, X__19 <lgl>, X__20 <chr>, X__21 <lgl>, X__22 <chr>,
## #   X__23 <lgl>, X__24 <chr>
```

Siin on mure, excelist on ka palju muud peale andmete. Peaksime mõned read algusest vahele jätma.

```
hdi_orig=read_excel("data/HDIdat.xls.xlsx", skip = 19)
head(hdi_orig)
```

```
## # A tibble: 6 x 25
##   `HDI Rank` Country `1980` X__1  `1985` X__2  `1990` X__3  `1995` X__4
##   <chr>      <chr>   <chr>  <lgl> <chr>  <lgl> <chr>  <lgl> <chr>  <lgl>
## 1 ..         Very h~ 0.766  NA    0.786  NA    0.810  NA    0.834  NA
## 2 ..         High h~ 0.614  NA    0.630  NA    0.648  NA    0.662  NA
## 3 ..         Medium~ 0.420  NA    0.450  NA    0.480  NA    0.517  NA
## 4 ..         Low hu~ 0.316  NA    0.334  NA    0.347  NA    0.363  NA
## 5 1          Norway  0.796  NA    0.819  NA    0.844  NA    0.876  NA
## 6 2          Austra~ 0.850  NA    0.859  NA    0.873  NA    0.889  NA
## # ... with 15 more variables: `2000` <chr>, X__5 <lgl>, `2005` <chr>,
## #   X__6 <lgl>, `2006` <chr>, X__7 <lgl>, `2007` <chr>, X__8 <lgl>,
## #   `2008` <chr>, X__9 <lgl>, `2009` <chr>, X__10 <lgl>, `2010` <chr>,
## #   X__11 <lgl>, `2011` <chr>
```

```
tail(hdi_orig,15)
```

```
## # A tibble: 15 x 25
##    `HDI Rank` Country `1980` X__1  `1985` X__2  `1990` X__3  `1995` X__4
##    <chr>      <chr>   <chr>  <lgl> <chr>  <lgl> <chr>  <lgl> <chr>  <lgl>
##  1 ..         Nauru   ..     NA    ..     NA    ..     NA    ..     NA
##  2 ..         Monaco  ..     NA    ..     NA    ..     NA    ..     NA
##  3 ..         Marsha~ ..     NA    ..     NA    ..     NA    ..     NA
##  4 ..         Korea ~ ..     NA    ..     NA    ..     NA    ..     NA
##  5 <NA>       <NA>    <NA>   NA    <NA>   NA    <NA>   NA    <NA>   NA
##  6 <NA>       <NA>    <NA>   NA    <NA>   NA    <NA>   NA    <NA>   NA
##  7 Footnotes  <NA>    <NA>   NA    <NA>   NA    <NA>   NA    <NA>   NA
##  8 <NA>       <NA>    <NA>   NA    <NA>   NA    <NA>   NA    <NA>   NA
##  9 <NA>       <NA>    <NA>   NA    <NA>   NA    <NA>   NA    <NA>   NA
## 10 Symbols    <NA>    <NA>   NA    <NA>   NA    <NA>   NA    <NA>   NA
## 11 ..         Data n~ <NA>   NA    <NA>   NA    <NA>   NA    <NA>   NA
## 12 (.)        Greate~ <NA>   NA    <NA>   NA    <NA>   NA    <NA>   NA
## 13 <          Less t~ <NA>   NA    <NA>   NA    <NA>   NA    <NA>   NA
## 14 -          Not ap~ <NA>   NA    <NA>   NA    <NA>   NA    <NA>   NA
## 15 T          Total   <NA>   NA    <NA>   NA    <NA>   NA    <NA>   NA
## # ... with 15 more variables: `2000` <chr>, X__5 <lgl>, `2005` <chr>,
## #   X__6 <lgl>, `2006` <chr>, X__7 <lgl>, `2007` <chr>, X__8 <lgl>,
## #   `2008` <chr>, X__9 <lgl>, `2009` <chr>, X__10 <lgl>, `2010` <chr>,
```

```
## #   X__11 <lgl>, `2011` <chr>
```

Ka lõpus on sama jama.

```
nrow(hdi_orig)
```

```
## [1] 209
```

```
hdi=hdi_orig[-c((nrow(hdi_orig)-10):nrow(hdi_orig)),]
tail(hdi)
```

```
## # A tibble: 6 x 25
##   `HDI Rank` Country `1980` X__1  `1985` X__2  `1990` X__3  `1995` X__4
##   <chr>      <chr>   <chr>  <lgl> <chr>  <lgl> <chr>  <lgl> <chr>  <lgl>
## 1 ..         Somalia ..     NA    ..     NA    ..     NA    ..     NA
## 2 ..         San Ma~ ..     NA    ..     NA    ..     NA    ..     NA
## 3 ..         Nauru   ..     NA    ..     NA    ..     NA    ..     NA
## 4 ..         Monaco  ..     NA    ..     NA    ..     NA    ..     NA
## 5 ..         Marsha~ ..     NA    ..     NA    ..     NA    ..     NA
## 6 ..         Korea ~ ..     NA    ..     NA    ..     NA    ..     NA
## # ... with 15 more variables: `2000` <chr>, X__5 <lgl>, `2005` <chr>,
## #   X__6 <lgl>, `2006` <chr>, X__7 <lgl>, `2007` <chr>, X__8 <lgl>,
## #   `2008` <chr>, X__9 <lgl>, `2009` <chr>, X__10 <lgl>, `2010` <chr>,
## #   X__11 <lgl>, `2011` <chr>
```

```
dim(hdi)
```

```
## [1] 198  25
```

```
#võime ka kohe ala ette anda
hdi=read_excel("data/HDIdat.xls.xlsx", range = "HDIdat!A20:Y218")
dim(hdi)
```

```
## [1] 198  25
```

```
head(hdi)
```

```
## # A tibble: 6 x 25
##   `HDI Rank` Country `1980` X__1  `1985` X__2  `1990` X__3  `1995` X__4
##   <chr>      <chr>   <chr>  <lgl> <chr>  <lgl> <chr>  <lgl> <chr>  <lgl>
## 1 ..         Very h~ 0.766  NA    0.786  NA    0.810  NA    0.834  NA
## 2 ..         High h~ 0.614  NA    0.630  NA    0.648  NA    0.662  NA
## 3 ..         Medium~ 0.420  NA    0.450  NA    0.480  NA    0.517  NA
## 4 ..         Low hu~ 0.316  NA    0.334  NA    0.347  NA    0.363  NA
## 5 1          Norway  0.796  NA    0.819  NA    0.844  NA    0.876  NA
## 6 2          Austra~ 0.850  NA    0.859  NA    0.873  NA    0.889  NA
## # ... with 15 more variables: `2000` <chr>, X__5 <lgl>, `2005` <chr>,
## #   X__6 <lgl>, `2006` <chr>, X__7 <lgl>, `2007` <chr>, X__8 <lgl>,
## #   `2008` <chr>, X__9 <lgl>, `2009` <chr>, X__10 <lgl>, `2010` <chr>,
## #   X__11 <lgl>, `2011` <chr>
```

```
tail(hdi)
```

```
## # A tibble: 6 x 25
##   `HDI Rank` Country `1980` X__1  `1985` X__2  `1990` X__3  `1995` X__4
##   <chr>      <chr>   <chr>  <lgl> <chr>  <lgl> <chr>  <lgl> <chr>  <lgl>
## 1 ..         Somalia ..     NA    ..     NA    ..     NA    ..     NA
## 2 ..         San Ma~ ..     NA    ..     NA    ..     NA    ..     NA
## 3 ..         Nauru   ..     NA    ..     NA    ..     NA    ..     NA
```

```
## 4 ..         Monaco  ..     NA    ..     NA    ..     NA    ..     NA
## 5 ..         Marsha~ ..     NA    ..     NA    ..     NA    ..     NA
## 6 ..         Korea ~ ..     NA    ..     NA    ..     NA    ..     NA
## # ... with 15 more variables: `2000` <chr>, X__5 <lgl>, `2005` <chr>,
## #   X__6 <lgl>, `2006` <chr>, X__7 <lgl>, `2007` <chr>, X__8 <lgl>,
## #   `2008` <chr>, X__9 <lgl>, `2009` <chr>, X__10 <lgl>, `2010` <chr>,
## #   X__11 <lgl>, `2011` <chr>
```

Mida veel tähele panna

Üldjuhul saab argumentidena ette anda mitmeid väärtusi. Olulisemad on neist seotud ka puuduvate väärtustega. Tühjad lahtrid saavad väärtuseks NA-not available. Kui puuduvat väärtust tähistab midagi muu, tuleb see ette öelda. Milline argument seda teeb, tuleb iga funktsiooni dokumentatsioonist vaadata (read_exceli puhul on selleks na, näiteks na="puuduv väärtus").

Muud allikad

Internet:

```
library(httr)
url='https://evs.nci.nih.gov/ftp1/CDISC/SDTM/SDTM%20Terminology.xls'
#tõmbab faili alla, teeb ajutise faili, leob sisse
GET(url, write_disk(tf <- tempfile(fileext = ".xls")))
df <- read_excel(tf, 2L)
```

Mis on eeltoodud näidise miinuseks?

Muud formaadid: https://www.statmethods.net/input/importingdata.html

Andmete salvestamine

RData formaat, suhteliselt efektiivne ja kiire.